Repérage de sens et désambiguïsation dans un contexte bilingue

Marianna Apidianaki

Repérage de sens et désambiguïsation dans un contexte bilingue

Marianna APIDIANAKI
Lattice, Université Paris 7, CNRS
ENS-1 rue Maurice Arnoux, F-92120, Montrouge
Marianna.Apidianaki@linguist.jussieu.fr

Résumé. Les besoins de désambiguïsation varient dans les différentes applications du Traitement Automatique des Langues (TAL). Dans cet article, nous proposons une méthode de désambiguïsation lexicale opératoire dans un contexte bilingue et, par conséquent, adéquate pour la désambiguïsation au sein d'applications relatives à la traduction. Il s'agit d'une méthode contextuelle, qui combine des informations de cooccurrence avec des informations traductionnelles venant d'un bitexte. L'objectif est l'établissement de correspondances de traduction au niveau sémantique entre les mots de deux langues. Cette méthode étend les conséquences de l'hypothèse contextuelle du sens dans un contexte bilingue, tout en admettant l'existence d'une relation de similarité sémantique entre les mots de deux langues en relation de traduction. La modélisation de ces correspondances de granularité fine permet la désambiguïsation lexicale de nouvelles occurrences des mots polysémiques de la langue source ainsi que la prédiction de la traduction la plus adéquate pour ces occurrences.

Abstract. Word Sense Disambiguation (WSD) needs vary greatly in different Natural Language Processing (NLP) applications. In this article, we propose a WSD method which operates in a bilingual context and is, thus, adequate for disambiguation in applications relative to translation. It is a contextual method which combines cooccurrence information with translation information found in a bitext. The goal is the establishment of translation correspondences at the sense level between the lexical items of two languages. This method extends the consequences of the contextual hypothesis in a bilingual framework assuming, at the same time, the existence of a semantic similarity relation between words of two languages being in a translation relation. The modelling of fine-grained correspondences allows for the disambiguation of new occurrences of the polysemous source language lexical items as well as for the prediction of the most adequate translation for those occurrences.

Mots clés : désambiguïsation contextuelle, similarité sémantique, substituabilité, traduction.

Keywords: contextual disambiguation, semantic similarity, substitutability, translation.

1. Désambiguïsation lexicale pour la traduction

La définition de la nature des sens, leur énumération et leur description constituent des questions centrales dans la problématique de la désambiguïsation lexicale, auxquelles une

réponse unanime est loin d'être trouvée. Les besoins concernant le degré de désambiguïsation ainsi que le type et le niveau des distinctions sémantiques varient dans le cadre de différentes applications du Traitement Automatique des Langues (TAL). Ainsi, les informations trouvées dans des ressources sémantiques prédéfinies s'avèrent souvent peu conformes aux besoins des applications particulières, et les méthodes de désambiguïsation sont parfois critiquées pour ne pas être liées à une application réelle.

Dans cet article, nous allons présenter une méthode originale de désambiguïsation lexicale qui opère dans un contexte bilingue et dont les résultats sont, par conséquent, utilisables dans des applications relatives à la traduction. Il s'agit d'une méthode de cooccurrences qui peut opérer sur les deux côtés d'un corpus parallèle (bitexte) : les contextes de la langue source (LS) et les contextes de la langue cible (LC). La combinaison d'informations contextuelles et traductionnelles permet le repérage de distinctions sémantiques au sein des mots polysémiques et l'établissement, entre les mots des deux langues, de correspondances au niveau sémantique exploitables dans des systèmes de traduction automatique ou assistée par ordinateur.

2. Principes de la méthode et hypothèses sous-jacentes

Les hypothèses théoriques sous-jacentes à cette méthode de désambiguïsation lexicale sont les suivantes :

1. l'hypothèse contextuelle du sens (Firth, 1957 ; Harris, 1985), d'après laquelle le sens des mots correspond à leurs usages dans les textes ;
2. l'hypothèse de l'existence d'une relation de similarité sémantique entre les mots de deux langues entretenant une relation de traduction dans des textes réels.

D'après la première hypothèse, l'analyse du contexte lexical (co-texte) entourant un mot dans des textes peut éclairer sa sémantique. Le contexte lexical a été exploité pour la désambiguïsation aussi bien dans des méthodes qui procèdent à la sélection du bon sens des mots à partir d'un dictionnaire dans un cadre monolingue (Lesk, 1986) et bilingue (Brun *et al*., 2001 ; Dufour, 1997), que dans des méthodes de désambiguïsation n'utilisant pas de ressources lexicales préalables. De telles méthodes, proposées dans un cadre monolingue, sont celles de Schütze (1998), de Véronis (2003) et de Pantel et Lin (2002) ; les deux premières exploitent les informations de cooccurrence des mots, tandis que la troisième met l'accent sur le contexte syntaxique. Dans un cadre de traduction, le contexte lexical des mots est exploré dans les méthodes de désambiguïsation proposées par Brown *et al*. (1991)[1], Kaji *et al*. (2003)[2] et Specia *et al*. (2006)[3], tandis que celle proposée par Dagan et Itai (1991) exploite le contexte syntaxique de la LC pour choisir le bon équivalent de traduction.

La deuxième hypothèse, citée plus haut, postule que, dans le cas de correspondances au niveau lexical au sein d'un corpus parallèle, le sens véhiculé par un équivalent de traduction est supposé similaire à celui du mot source qu'il traduit. Par conséquent, les équivalents de traduction possibles des mots polysémiques de la LS sont censés traduire les différents sens

[1] La méthode utilise des questions binaires pour choisir entre deux sens d'un mot.

[2] La méthode exploite des corpus comparables et utilise des informations sur l'alignement translinguistique de paires de mots liés. Les sens sont décrits par un ou plusieurs équivalents, dont le regroupement se base sur la similarité distributionnelle dans les deux langues. Cette méthode ne prend pas en compte les ambiguïtés parallèles entre les mots des deux langues, et elle présuppose que chaque équivalent traduit uniquement un sens du mot polysémique.

[3] La méthode proposée par Specia *et al*. prend en compte des informations de cooccurrence lexicale dans la LC acquises à l'aide de requêtes effectuées sur le Web concernant des fragments de texte de la LC.

de ces mots dans la LC, sens reflétés aussi dans le contexte lexical de la LS. La nouveauté de notre approche consiste justement au repérage automatique des sens des mots polysémiques par projection des informations de cooccurrence d'un côté du bitexte à l'autre, sans recours à une ressource lexicale préalable. L'analyse sémantique qui en résulte concerne tant les mots de la LS que les équivalents, et les résultats sont directement exploitables pour la désambiguïsation et la sélection lexicale dans la traduction.

La correspondance sémantique entre deux unités lexicales en relation de traduction peut être mise en évidence et servir à la modélisation de correspondances sémantiques au sein d'un système automatique. La correspondance à laquelle nous faisons face avant la désambiguïsation d'un mot polysémique de la LS se situe au niveau lexical, où le mot en question correspond à plusieurs équivalents dans la LC. Le but est de raffiner cette relation et de représenter les liens entre les mots des deux langues à un niveau d'analyse plus élevé. Sur la base des hypothèses précitées, nous acceptons que les informations venant du co-texte des occurrences de l'unité source qui sont traduites par un équivalent précis dans le corpus éclairent tant le(s) sens véhiculé(s) par ces occurrences que celui(ceux) de l'équivalent de traduction. Le co-texte du mot source par rapport à un équivalent précis correspond aux mots de contenu (noms, adjectifs et verbes) qui cooccurrent avec le mot dans les segments de traduction où il est traduit par cet équivalent[4].

Dans une méthode contextuelle monolingue de désambiguïsation, la comparaison des contextes (lexicaux ou grammaticaux) des occurrences du mot polysémique permet leur clusterisation en fonction de leur similarité et les clusters résultants sont censés illustrer les différents sens du mot. Ici, au lieu de comparer entre eux et de clusteriser les contextes dans lesquels un mot polysémique apparaît, nous allons comparer entre eux et clusteriser des ensembles de contextes correspondant à chacun de ses équivalents. Dans le paragraphe suivant, nous allons décrire comment les ensembles en question sont construits.

3. Prétraitement du corpus

Le corpus utilisé dans ce travail pour l'apprentissage est un bitexte anglais-grec de 4 000 000 mots (Gavrilidou *et al.,* 2004), lemmatisé, morphosyntaxiquement étiqueté et aligné au niveau des phrases et au niveau des mots (Simard, Langlais, 2003). La source principale des textes est le *Journal de l'Union Européenne* (domaines : droit [42 % des textes du corpus], santé [24 %], éducation [21 %]), mais il y a aussi des textes venant de l'Office National Hellénique du Tourisme (11 %), ainsi qu'un petit nombre de textes scientifiques sur l'environnement (2 %). Pour chaque mot polysémique étudié, un sous-corpus a été créé, qui contient les segments de traduction dans lesquels le mot source occurre[5]. Le choix des segments de traduction en tant que contexte est dicté par notre objectif d'exploration de l'influence du co-texte proche des unités source sur la désambiguïsation et le transfert lexical. Les segments constituant le sous-corpus d'un mot polysémique ont été regroupés en fonction de ses équivalents de traduction. Ainsi des ensembles de phrases correspondant à chacun des équivalents sont créés dans les deux côtés du bitexte, comme cela est décrit dans la figure 1.

Cette figure illustre le sous-corpus d'un mot source *m*, qui est traduit dans le corpus par trois équivalents différents : *a*, *b* et *c*. Dans la partie gauche de la figure, nous avons les phrases de la LS et, à droite, leurs traductions dans la LC, qui se trouvent dans les mêmes segments de

[4] Tous les calculs opèrent sur les lemmes (*types*) auxquels les mots des contextes (*tokens*) ont été ramenés.
[5] Un segment peut contenir de 0 à 2 phrases par langue. Par exemple, un alignement *2:1* met en correspondance 2 phrases du texte de la LS avec 1 phrase du texte de la LC, à l'intérieur d'un segment.

traduction, comme cela a été déterminé par le processus d'alignement des phrases.

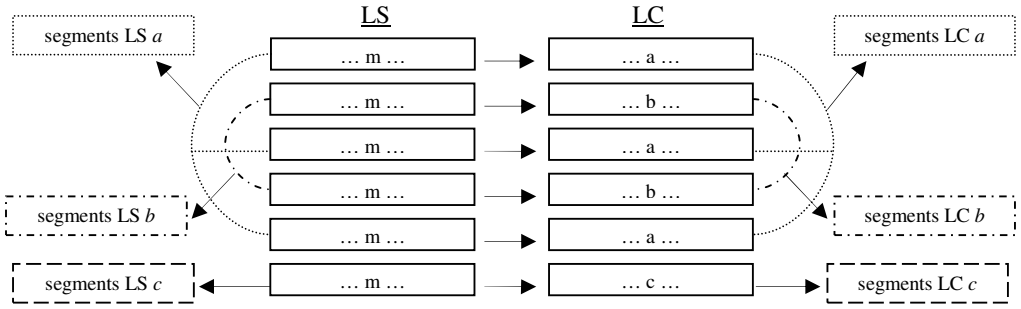

Figure 1 : Regroupement des segments de traduction en fonction des équivalents

Les segments sont regroupés en fonction des équivalents. Ainsi, nous avons un ensemble de phrases dans la LS correspondant aux occurrences de *m* traduites par *a* et un ensemble de phrases dans la LC correspondant à ces traductions et contenant, bien évidemment, le mot *a*. Nous procédons de la même manière pour les autres équivalents (*b* et *c*) en constituant les groupes de phrases respectifs dans les deux langues[6]. Ces ensembles de segments constituent l'entrée de la méthode de similarité sémantique qui sera présentée par la suite.

4. Méthode d'estimation de similarité sémantique

4.1 Description et présupposés

Le calcul de similarité sémantique est appliqué sur les ensembles de segments correspondant aux équivalents de traduction d'un mot polysémique. Il peut porter sur les contextes du mot source aussi bien que sur les contextes des équivalents. Sur la base de l'hypothèse contextuelle du sens, un grand degré de similarité des contextes de la LS indique l'homogénéité sémantique du mot source, tandis que leur dissimilarité constitue un indice de l'existence de distinctions sémantiques au sein du mot en question. D'après la deuxième hypothèse du départ, qui concerne l'existence d'une correspondance d'ordre sémantique entre les mots de deux langues entretenant une relation de traduction, la similarité plus ou moins grande des contextes de la LS correspondant à des équivalents différents indique le degré de similarité sémantique des équivalents en question. Ainsi, les résultats du calcul de similarité au sein de la LS permettent la clusterisation (ou la distinction) des équivalents en fonction de leur similarité (ou dissimilarité) sémantique.

Les équivalents clusterisés sur la base de leur similarité sémantique sont censés traduire le même sens du mot polysémique[7]. Dans une approche contextuelle de la similarité sémantique, les mots similaires sont considérés comme substituables au sein des contextes qui induisent leur relation (Miller et Charles, 1991). Ainsi, il est possible d'émettre l'hypothèse que lorsque le calcul porte sur les contextes de la LS, les équivalents clusterisés sont substituables en tant que traductions pour les occurrences du mot source trouvées dans les contextes révélant leur similarité.

La projection des résultats de la clusterisation sur le mot source permet donc le repérage des sens véhiculés par le mot. Ainsi la notion de similarité sémantique basée sur la similarité contextuelle, et hautement utilisée dans le cadre de la désambiguïsation monolingue, est

[6] Dans la figure, nous décrivons ces ensembles de phrases comme « segments LS *équivalent* » et « segments LC *équivalent* ».

[7] En revanche, la distinction des équivalents signale leur dissimilarité sémantique, c'est-à-dire qu'ils traduisent des sens différents du mot source.

reprise ici afin d'émettre des jugements sur la similarité des équivalents de traduction qui peuvent, par la suite, être projetés sur les unités polysémiques source. Un avantage de l'utilisation de cette méthode par rapport à l'application d'une méthode monolingue de désambiguïsation (Apidianaki, 2006) est que les distinctions proposées sont beaucoup plus pertinentes pour la traduction[8].

4.2 La mesure de similarité sémantique

La mesure utilisée pour estimer la similarité sémantique entre les équivalents de traduction est une variation de la mesure de Jaccard pondérée (JP), proposée par Grefenstette pour la création de thesaurus sémantiques dans un contexte monolingue (1994 : 48-50)[9]. Dans notre travail, les informations utilisées pour le calcul concernent la cooccurrence des mots à l'intérieur des segments de traduction. Le contexte de la LS par rapport à un équivalent est constitué par les mots de contenu cooccurrant avec le mot source dans les segments où il est traduit par cet équivalent précis. D'autre part, les traits sur lequel porte le calcul dans la LC sont les mots de contenu du co-texte de l'équivalent dans les traductions. Pour chaque trait nous calculons son **poids global** (global weight : *gw*) :

$$gw(trait_j) = 1 - \sum_i \frac{p_{ij} \log(p_{ij})}{nrels} \qquad \text{où} \quad p_{ij} = \frac{fréquence\,absolue\,du\,trait_j\,avec\,l'équiv_i}{nombre\,total\,de\,traits\,pour\,l'équiv_i}$$

$$\text{et} \quad nrels = le\,nombre\,total\,de\,relations\,extraites\,du\,corpus\,pour\,le\,trait_j$$

son **poids local** (local weight : *lw*) : $lw(équiv_i, trait_j) = \log(fréquence\,du\,trait_j\,avec\,l'équiv_i)$

et son **poids total** (whole weight : *W*) : $W = gw * lw$

Les éléments du contexte qui nous servent comme traits sont donc pondérés en fonction de leur dispersion dans les textes (*gw*) et de leur fréquence d'occurrence avec chaque équivalent précis (*lw*). Le poids global d'un trait est calculé en prenant en compte la somme de la probabilité d'occurrence du trait avec chacun des équivalents (p_{ij}), ainsi que le nombre total d'équivalents avec lesquels ce trait occurre (*nrels*). Le poids local est calculé sur la base de la fréquence d'apparition du trait avec un équivalent précis. Ainsi le coefficient pondéré nous permet d'attribuer une importance aux traits, qui est proportionnelle à leur pertinence pour l'estimation de la similarité entre les équivalents. Le Jaccard entre deux équivalents *m* et *n* est calculé par la formule suivante :

$$JP(équiv_m, équiv_n) = \frac{\sum_j \min(W(équiv_m, trait_j)\, W(équiv_n, trait_j))}{\sum_j \max(W(équiv_m, trait_j)\, W(équiv_n, trait_j))}$$

4.3 Analyse des résultats

Le processus a été appliqué sur dix mots polysémiques anglais[10]. Nous allons présenter ici un exemple illustrant le fonctionnement de la méthode et les résultats que nous pouvons obtenir. Cet exemple concerne le mot *movement* qui a neuf équivalents de traduction dans le corpus[11]: *κυκλοφορία* (251), *διακίνηση* (38), *κίνηση* (28), *μετακίνηση* (19), *κίνημα* (11), *κινητικότητα*

[8] Les sens proposés par l'application d'une méthode de désambiguïsation qui ne prend pas en compte les équivalents dès le début sont très nombreux, et il est difficile de créer des correspondances de traduction satisfaisantes au niveau sémantique. L'utilisation des équivalents comme indices pour le fusionnement des sens risque de nous faire entrer dans un cercle vicieux à cause de la polysémie propre aux équivalents mêmes.

[9] La similarité entre les mots est estimée sur la base de leurs contextes syntaxiques partagés.

[10] *Plant*, *movement*, *occupation*, *communication*, *treatment*, *passage*, *power*, *competence*, *facility*, *paper*.

[11] Entre parenthèses, nous donnons la fréquence avec laquelle l'équivalent traduit le mot source dans le corpus.

(6), *προσπάθεια* (1), *τάση* (1), *βήμα* (1). Les résultats du calcul de similarité entre ces équivalents sont décrits dans le tableau 1[12].

Paires d'équivalents		Contextes anglais	Contextes grecs
μετακίνηση	διακίνηση	**0,11**	**0,125**
κίνηση	διακίνηση	**0,099**	**0,13**
μετακίνηση	κίνηση	**0,087**	**0,141**
κυκλοφορία	διακίνηση	**0,078**	**0,091**
κίνηση	κυκλοφορία	**0,063**	**0,077**
κίνηση	κίνημα	**0,046**	0,052
	Moyenne	**0,043**	**0,062**

Tableau 1 : Relations de similarité entre les équivalents de *movement*

Les éléments des paires qui apparaissent au début de la liste sont censés entretenir des relations sémantiques plus fortes que ceux qui apparaissent vers la fin. L'analyse des résultats obtenus pour tous les mots étudiés a démontré que la moyenne des scores associés aux paires d'équivalents[13] pourrait constituer une sorte de seuil au-dessous duquel les relations trouvées ne sont pas pertinentes. Les relations repérées permettent la clusterisation des équivalents : les éléments inclus dans un cluster sont censés traduire le même sens du mot source tandis qu'un élément qui appartient à plusieurs clusters est supposé traduire des sens différents.

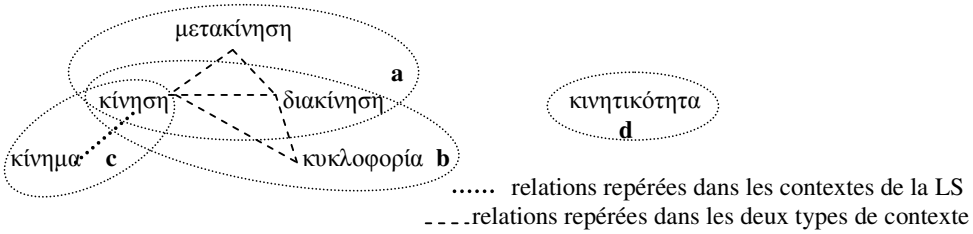

Figure 2 : Clusterisation des équivalents de *movement*

Les relations entre les équivalents sont induites soit par les contextes de la LS, soit par ceux de la LC, soit par les deux types de contexte ; ceci est illustré dans la figure 2 par l'utilisation de lignes différentes. D'après les résultats obtenus jusqu'à maintenant, les relations repérées seulement dans les contextes des traductions ne sont pas pertinentes, contrairement aux relations repérées dans les contextes des textes originaux ou dans les deux types de contextes.

La projection des clusters d'équivalents sur le mot polysémique source induit trois sens au sein du mot *movement*. Les deux premiers sens sont illustrés par les clusters *a* et *b* (*μετακίνηση-κίνηση-διακίνηση* et *διακίνηση-κίνηση-κυκλοφορία*) et correspondent aux usages de *movement* qui décrivent les notions de mobilité, de circulation et de transfert. Les sens décrits par ces deux clusters pourraient théoriquement être regroupés dans un sens plus large. Néanmoins, la distinction induite par les clusters séparés peut être expliquée comme reflétant des contraintes d'utilisation et de substitution des équivalents au sein des contextes. Le cluster *c* (*κίνημα-κίνηση*), décrit le sens métaphorique de *movement* (par ex. des mouvements sociaux et autres). Ce sens est le plus souvent traduit par *κίνημα*, mais aussi parfois par *κίνηση*. *Κίνηση* est un mot polysémique qui peut véhiculer, d'une part, le sens de mobilité et de déplacement physique et, d'autre part, le sens de mouvement social. L'isolation de certains équivalents, comme dans le cas de *κινητικότητα*, peut être due aussi bien à des raisons sémantiques qu'à leur basse fréquence d'occurrence dans le corpus, ce qui fait qu'il n'y a pas assez

[12] Paires faiblement liées (avec des scores inférieurs à la moyenne) : *μετακίνηση-κυκλοφορία* (0,035-0,05) / *-κινητικότητα* (0,039-0,009) / *-κίνημα* (0,009-0,044), *κινητικότητα-διακίνηση* (0,038-0,036) / *-κυκλοφορία* (0,022-0,023) / *-κίνημα* (0-0,045) / *-κίνηση* (0-0,029), *κίνημα-διακίνηση* (0,019-0,06) / *-κυκλοφορία* (0,008-0,018).

[13] La moyenne des scores attribués à toutes les paires (mêmes aux paires d'équivalents faiblement liés).

d'informations contextuelles pour les rapprocher des autres. Quand la fréquence d'occurrence d'un équivalent est grande par rapport à celle des autres équivalents et qu'il reste, malgré cela, isolé, nous pouvons supposer l'existence d'une différence sémantique. Dans le cas de *κινητικότητα*, son isolation est due à sa basse fréquence dans le corpus (6) et à son contexte très restreint, qui fait qu'il n'y a pas assez d'informations contextuelles sur cet équivalent qui pourraient le « rapprocher » des autres.

Les relations entre les équivalents contenus dans un cluster sont modélisées à l'aide des éléments contextuels qui les mettent en évidence, c'est-à-dire par les traits qui sont communs aux équivalents en question et que nous pourrions appeler leurs *co-textes assimilateurs* (Fuchs, 1994 :134-141). Ainsi chaque paire d'équivalents est caractérisée par l'ensemble de traits de leurs co-textes assimilateurs et par un ensemble de traits correspondant à chacun des équivalents, qui contient aussi bien leurs traits communs que les traits de leurs co-textes *dissimilateurs*, qui différencient l'un équivalent de l'autre. Il se peut que les co-textes dissimilateurs caractérisant un équivalent décrivent un sens véhiculé seulement par celui-ci (et non par l'autre membre de la paire). Le regroupement des traits dissimilateurs et assimilateurs au sein de ces ensembles provoque une perte d'informations relatives à ce type de distinctions qui n'a pas d'impact négatif important sur le processus de prédiction de traduction et qui améliore même les résultats du point de vue qualitatif. Le regroupement permet la sélection correcte de l'un des équivalents d'une paire, non pas seulement si celui-ci véhicule un sens bien distinct, mais aussi lorsqu'il est plus adéquat que l'autre équivalent de la paire pour traduire la nouvelle occurrence du mot source. En revanche, l'utilisation de correspondances de ce type dans le cadre d'une autre application, comme la recherche d'informations multilingues, nécessiterait probablement une modélisation plus fine des sens véhiculés par chacun des équivalents.

5. Évaluation

Le corpus de test utilisé pour l'évaluation de la méthode est différent du corpus d'apprentissage. Il s'agit de la partie anglais-grec du corpus parallèle EUROPARL (Koehn, 2003). Etant aligné au niveau des phrases, nous avons pu extraire à partir de ce corpus des segments de traduction contenant, d'une part, les phrases en anglais où les mots polysémiques occurrent et, d'autre part, leurs traductions en grec. L'évaluation de la méthode consiste à l'utilisation des correspondances établies entre les mots polysémiques et leurs équivalents de traduction pendant l'étape précédente pour (a) la désambiguïsation des nouvelles occurrences des mots polysémiques et (b) la prédiction des traductions les plus adéquates pour ces occurrences. Pour chacun des équivalents des mots polysémiques trouvés dans le corpus de test, nous avons retenu un ensemble de segments choisis au hasard. Par exemple, le mot *movement* est traduit dans EUROPARL par tous les équivalents clusterisés (voir figure 2) et nous avons retenu dix segments correspondant à chacun de ces équivalents.

Etant donné que l'apprentissage a opéré sur les formes des mots de catégories grammaticales précises et que les correspondances ont été modélisées à l'aide de celles-ci, il a fallu lemmatiser et étiqueter morphosyntaxiquement les nouveaux contextes pour rendre possible leur comparaison avec les informations retenues[14]. Voici un exemple d'un segment de traduction pour le mot *movement* :

[14] L'étiquetage morphosyntaxique et la lemmatisation ont été effectués à l'aide de l'étiqueteur TreeTagger (Schmid, 1994). Nous rappelons que seulement les noms, les adjectifs et les verbes des contextes sont retenus.

{I am therefore delighted that steps are finally being taken which will allow the theoretical freedom of movement of persons to be translated into practice, albeit still far from perfect practice! -- *Επομένως είμαι ικανοποιημένος που επιτέλους λαμβάνονται μέτρα τα οποία θα επιτρέψουν να γίνει πράξη η αρχή της ελεύθερης* ***κυκλοφορίας*** *των προσώπων, έστω και αν τα μέτρα αυτά είναι ακόμα, κατά τη γνώμη μου, ανεπαρκέστατα!*}

Le contexte de la nouvelle occurrence de *movement* est représenté de la manière suivante : {*be* (VBP), *delight* (VVN), *step* (NNS), *be* (VBG), *take* (VVN), *allow* (VV), *theoretical* (JJ), *freedom* (NN), *person* (NNS), *translate* (VVN), *practice* (NN), …}. La comparaison de cet ensemble d'éléments avec les traits qui caractérisent les correspondances entre *movement* et ses équivalents permet la sélection de l'équivalent le plus adéquat pour la nouvelle occurrence du mot. La prédiction de traduction avec le plus grand score pour cette occurrence de *movement* est la paire d'équivalents : *διακίνηση-κυκλοφορία* (score : 2,951).

La proposition de paires d'équivalents pour un mot constitue l'un des points forts de la méthode dans le sens où elle profite bien des informations paradigmatiques relatives à la similarité sémantique entre les équivalents, qui enrichissent les correspondances établies entre les mots des deux langues. La proposition d'une paire d'équivalents signifie qu'ils sont interchangeables au sein du nouveau contexte. Il faut souligner le rôle important des scores attribués aux nouveaux contextes par rapport aux correspondances établies ; ces scores sont calculés sur la base des poids attribués aux traits qui caractérisent les équivalents clusterisés. Ainsi, une proposition est faite non pas sur la base du nombre des traits communs entre le nouveau contexte et les correspondances modélisées, mais en fonction de la pertinence des traits du nouveau contexte par rapport aux éléments qui forment les correspondances.

Pour évaluer les résultats du point de vue quantitatif, nous utilisons les mesures de rappel et de précision ; le rappel correspond au rapport du nombre de prédictions correctes au nombre de traductions de référence et la précision au rapport du nombre de prédictions correctes au nombre de propositions faites par le système. Nous considérons comme prédiction correcte la proposition de l'équivalent qui traduit la nouvelle occurrence du mot dans le corpus de test – dans notre exemple il s'agit de l'équivalent *κυκλοφορία* – ainsi que les propositions d'équivalents qui retiennent une relation de similarité sémantique avec lui et donc se trouvent dans le même cluster[15]. La prise en compte des relations paradigmatiques établies entre les équivalents et décrites à l'aide des clusters apporte une amélioration des résultats de l'évaluation par rapport aux résultats obtenus en considérant comme correctes seulement les propositions correspondant aux équivalents trouvés dans le corpus de référence. Dans le premier cas, la précision pour les dix mots étudiés est de 71,91 % et le rappel de 68,26 %, tandis que, dans le deuxième cas, la précision est de 38,48 % et le rappel de 36,53 %. Nous remarquons aussi l'influence des limitations du corpus sur les résultats ; ceux-ci se détériorent clairement dans le cas d'équivalents qui ont une fréquence d'occurrence inférieure à dix dans le corpus d'apprentissage. Les correspondances modélisées entre le mot source et ces équivalents ne contiennent pas d'informations suffisantes sur leurs conditions d'utilisation afin qu'ils soient sélectionnés pour traduire une nouvelle occurrence du mot. Si les résultats pour ces équivalents ne sont pas pris en compte, la précision est de 80,33 % et le rappel de 76,8 % en tenant compte de la clusterisation versus 48,11 % et 46 % dans le cas inverse.

[15] Par exemple, la proposition de l'équivalent *διακίνηση* pour une occurrence de *movement* traduite dans le corpus de test par *κυκλοφορία*. Cette proposition est considérée comme correcte étant donné que les deux équivalents sont liés et qu'ils sont considérés comme adéquats et interchangeables au sein du nouveau contexte.

6. Discussion et perspectives

La méthode présentée dans cet article ne se base sur aucune ressource sémantique prédéfinie, et les sens sont mis en évidence sur la base des informations contenues dans le corpus d'apprentissage. Il serait donc souhaitable de procéder à une validation des sens proposés et des relations sémantiques établies entre les équivalents. Pour ce faire, nous avons comparé les résultats obtenus par la méthode des cooccurrences avec ceux d'une autre méthode automatique de désambiguïsation ; il s'agit de la méthode des Miroirs Sémantiques proposée par Dyvik (1998, 2003) qui se sert des informations sur les relations de traduction entre les mots d'un bitexte afin d'analyser leur sémantique et de construire des thesaurus sémantiques. En raison des contraintes d'espace, nous ne pouvons pas présenter ici en détail les résultats de l'application de cette méthode sur notre corpus. Cependant, nous pouvons signaler que les distinctions sémantiques proposées par cette méthode sont assez similaires aux distinctions proposées par la méthode contextuelle. Etant donné que les Miroirs Sémantiques utilisent des informations de nature très différente[16], la similarité entre les résultats des deux méthodes constitue une sorte de validation des résultats obtenus par la méthode contextuelle et apporte un bon degré de confiance quant aux distinctions sémantiques proposées.

L'application à notre corpus de la méthode des Miroirs Sémantiques permet aussi de remédier à une limitation de la méthode contextuelle de désambiguïsation. Celle-ci rend possible le repérage des relations de similarité sémantique entre les équivalents et leur clusterisation, mais elle ne permet pas l'analyse de la nature des relations entre les équivalents clusterisés. Les différents équivalents de traduction d'un mot peuvent entretenir des relations de (quasi)-synonymie, d'hyponymie, d'hyperonymie ou, même, de causalité. La méthode des Miroirs Sémantiques offre la possibilité d'analyse de la nature des relations sémantiques existant entre les équivalents, qui sont décrites au sein des entrées correspondantes du thesaurus.

Compte tenu, d'une part, du caractère automatique de la méthode de désambiguïsation proposée dans cet article et, d'autre part, des résultats encourageants de l'évaluation, nous estimons que cette méthode mérite d'être testée à grande échelle. La manière dont l'évaluation est menée ici (par prise en compte d'un nombre fixe et plus ou moins égal de nouvelles occurrences par candidat de traduction) ne permet pas de comparer les résultats avec une méthode « baseline », qui concernerait, par exemple, la sélection de l'équivalent (et du sens) le plus fréquent pour toutes les occurrences[17]. Par la suite, nous procéderons à une évaluation plus globale de la méthode sur l'ensemble du corpus de test, qui nous permettra aussi d'avoir des résultats significatifs à l'aide de la méthode « baseline », comparables avec les résultats de notre méthode.

Références

APIDIANAKI M. (2006) Traitement de la polysémie lexicale dans un but de traduction. Actes de *TALN 2006*, 10-13 avril, Leuven, Belgique, vol. 1, 53-62.

BROWN P. F., DELLA PIETRA S. A., DELLA PIETRA V. J., MERCER R. L. (1991) Word-sense disambiguation using statistical methods. Actes de *29th Annual Meeting of the Association for Computational Linguistics,* 264-270.

[16] Les informations contextuelles ne sont pas prises en compte.

[17] Nous estimons que la comparaison avec une méthode « baseline » de ce type serait beaucoup plus intéressante que la comparaison avec une méthode effectuant le choix aléatoire entre les alternatives.

BRUN C., JACQUEMIN B., SEGOND F. (2001) Exploitation de dictionnaires électroniques pour la désambiguïsation sémantique lexicale, TAL, vol. 42(3), 667-690.

DAGAN I., ITAI A. (1991) Word Sense Disambiguation Using a Second Language Monolingual Corpus, Computational Linguistics, vol. 20(4), 563-596.

DUFOUR N. (1997) DEFIDIC, a lexical database for computerized translation selection, *RISHH* vol. 33, Liège, 79-111.

DYVIK H. (1998) A translational basis for semantics. *Corpora and Cross-Linguistic Research : Theory, Method and Case Studies*, Johansson S., Oksefjell S. (éds.), 51-86.

DYVIK H. (2003) Translations as a semantic knowledge source. (brouillon) URL : http://www.hf.uib.no/i/LiLi/SLF/ans/Dyvik/transknow.pdf

FUCHS C. (1994) *Paraphrase et énonciation*. Paris : Editions Ophrys.

GAVRILIDOU M., LABROPOULOU P., DESIPRI E., GIOULI V., ANTONOPOULOS V., PIPERIDIS S., (2004) Building parallel corpora for eContent professionals. Actes de *MLR 2004, PostCOLING Workshop on Multilingual Linguistic Resources,* Genève, 28 août.

GREFENSTETTE G. (1994) *Explorations in Automatic Thesaurus Discovery*. Boston/Dordrecht/London : Kluwer Academic Publishers.

JOHANSSON S., OKSEFJELL S. (éds.) (1998) *Corpora and Cross-Linguistic Research : Theory, Method and Case Studies*. Amsterdam/Atlanta : Rodopi.

KAJI H. (2003) Word Sense Acquisition from Bilingual Comparable Corpora. Actes de *HLT-NAACL*, Edmonton, mai-juin, 32-39.

KOEHN P. (2003) *Europarl : a Multilingual Corpus for Evaluation of Machine Translation*. (brouillon) URL : http://www.iccs.inf.ed.ac.uk/~pkoehn/publications/europarl.pdf

MILLER G. A., CHARLES W.G. (1991) Contextual Correlates of Semantic Similarity. *Language and Cognitive Processes*, vol. 6(1), 1-28.

PANTEL P., LIN D. (2002) Discovering Word Senses from Text. Actes de *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Edmonton, 613-619.

SCHÜTZE H. (1998) Automatic Word Sense Discrimination. *Computational Linguistics,* vol. 24(1), 97-123.

SCHMID H. (1994) Probabilistic Part-of-Speech Tagging Using Decision Trees. Actes de *International Conference on New Methods in Language Processing*, Manchester, 44-49.

SIMARD M., LANGLAIS P. (2003) Statistical Translation Alignment with Compositionality Constraints. Actes de *HLT-NAACL Workshop : Building and Using Parallel Texts : Data-Driven Machine Translation and Beyond,* Edmonton, Canada, May 31, 19-22.

SPECIA L., VOLPE NUNES M. DAS GRAÇAS, STEVENSON M. (2006) Translation Context Sensitive WSD, Actes de *EAMT-2006 : 11th Annual Conference of the European Association for Machine Translation*, 19-20 juin, Oslo, 227-232.

VERONIS J. (2003) Hyperlex : cartographie lexicale pour la recherche d'informations. Actes de *TALN 2003*, Batz-sur-mer, 11-14 juin, 265-274.

VICKREY D., BIEWALD L., TEYSSIER M., KOLLER D. (2005) Word-Sense Disambiguation for Machine Translation, Actes de *HLT/EMNLP*, Vancouver, BC, 771-778.